“仁者”还是“智者”：第三方惩罚对惩罚者声誉的影响*

陈思静　徐烨超

（浙江科技学院经济与管理学院，杭州，310023）

摘要 第三方惩罚会对惩罚者的声誉产生显著影响，然而就影响的方向而言，现有文献给出了不同答案。上述问题的一个潜在原因是先前研究未能区分声誉的不同维度以及惩罚的不同动机与形式。通过将温暖-能力双维度结构引入惩罚者的声誉，实验结果显示，第三方惩罚从总体上降低了人们对惩罚者在温暖维度上的评价而提高了对其在能力维度上的评价。调节效应分析表明，动机被归因为集体聚焦的惩罚进一步提升了其对能力的正面作用而减缓了对温暖的负面作用，并且惩罚者的合作水平越高，其动机被归因为集体聚焦的程度也越高。针对不同惩罚形式的进一步分析显示，当惩罚动机被归因为个体聚焦时，经济惩罚对温暖的负面作用显著高于社会惩罚，而在集体聚焦的归因下经济惩罚对能力的正面作用显著低于社会惩罚。

关键词 第三方惩罚；社会规范；惩罚动机；声誉；经济惩罚；社会惩罚

分类号 B849: C91

1 前言

在社会科学文献中，合作通常被定义为个体付出成本而使他人受益的行为（Rand, 2016），合作是人类社会大量问题得以解决的关键所在（Bear & Rand, 2016），为此我们发展出了合作的规范（de Kwaadsteniet et al., 2007; Fehr & Schurtenberger, 2018）。然而对合作规范的遵守并非自然而然之事，因为个体倾向于追求私利，而这通常导致公共产品的供给不足和社会运行效率的损失（de Kwaadsteniet et al., 2019）。那么，非亲缘个体间的大规模合作是如何得以维系的？Fehr 和 Gächter（2002）的第三方惩罚理论为此提供了部分解释，该理论认为某些个体具有惩罚规范破坏者的先天倾向，只要存在足够数量的此类个体，那么群体成员间的合作关系就能得以维系（Carpenter et al., 2009）。但由于第三方惩罚的成本（花费的金钱、时间、精力以及可能受到的潜在报复）由惩罚者承担，而收益却由群体成员共享，第三方惩罚引发了二阶社会两难问题（second-order social dilemma）（Colman, 2006; Hauert et al., 2007）：相对于惩罚性合作者（下简称惩罚者），只合作但不惩罚的个体就是一个二阶搭便车者（second-order free rider）。由于惩罚成本由惩罚者承担，二阶搭便车者的演化适应度必然

收稿日期：2020 年 3 月 2 日

* 国家自然科学基金项目（71701185）、浙江省软科学项目（2020C35020）资助。

通信作者：陈思静，E-mail: chensijing@zust.edu.cn

高于惩罚者（谢晓非等, 2017; Hu et al., 2016），这又提出了一个新的问题：惩罚者是如何从演化中胜出的？

一种广受关注的观点是第三方惩罚能为惩罚者带来积极的声誉（Barclay, 2006; Barclay & Kiyonari, 2014），而积极的声誉能带来相应的收益，比如惩罚者在未来人际互动中得到他人帮助或奖励的概率得以提升（Santos et al., 2010），或向外界传达了惩罚者拥有良好品质的可靠信号（Jordan et al., 2016）。如果这种收益能超过惩罚成本，那么惩罚者就能在演化中得到选择。上述观点主要基于间接互惠理论（indirect reciprocity theory）或高成本信号理论（costly signaling theory），一方面为第三方惩罚的演化提供了理论解释，并获得了一定数量的研究支持（e.g., Jordan & Rand, 2019; Kurzban et al., 2007），另一方面，上述观点的前提条件是惩罚者的声誉必然是积极的。然而，有越来越多的证据显示，惩罚者的声誉未必是积极的（Bornstein & Weisel, 2010），甚至很有可能是负面的（de Kwaadsteniet et al., 2019; Ozono & Watabe, 2012），而且惩罚也未必能提高惩罚者得到他人帮助的概率（Kiyonari & Barclay, 2008）。这就意味着第三方惩罚与惩罚者声誉之间的关系可能比我们预想的要更为复杂，因此我们需要更深入地检视惩罚者的声誉机制，从而才能有效探讨声誉能否充分解释第三方惩罚的演化优势。

Rand 和 Nowak（2013）注意到，探讨合作的主流演化理论往往将个体简化为不具备动机的行动者（agent），而完全忽视了心理动机的重要性。这可能是因为目前有关合作与惩罚的文献主要来自经济学、生物学和博弈论等领域（陈思静, 杨莎莎, 印刷中），而心理学视角的缺席导致我们在很大程度上忽略了动机在惩罚者声誉机制中的作用。事实上，人们总是依据动机来对他人行为做出道德判断进而影响后续的人际互动（Bigman & Tamir, 2016），这意味着同样的行为在不同动机归因下会对人际关系产生截然不同的效果。就惩罚而言，Fehr 和 Rockenbach（2003）以及刘国芳和辛自强（2014）的研究证实了动机归因显著影响了惩罚对合作的效果：只有当惩罚者的动机被归因为利他时，第三方惩罚才能促进受罚者的合作水平，反之，惩罚则抑制了受罚者的合作行为。一个合理的推测是惩罚动机对惩罚者的声誉也具有相似的作用机制，即只有动机合理的第三方惩罚才能提升惩罚者的声誉。基于上述推理，本文提出研究问题 1：人们对第三方惩罚动机的归因是否会显著影响惩罚者的声誉？

其次，先前研究者倾向于将声誉当作一个单维度概念（uni-dimensional variable），而忽略了其不同维度，这导致先前研究中惩罚者声誉要么是全然积极的，要么就是全然消极的（de Kwaadsteniet et al., 2019）。正如 Beersma 和 van Kleef（2011）指出，声誉本质上是个体对他人的感知或评价，而在相应文献中，一个重要发现是人们通常运用两个基本维度来形成对他

人的评价（Fiske et al., 2007）：温暖（warmth）和能力（competence）。温暖指的是个体在与他人互动中表现出来的良善特质，如值得信赖；而能力指的是一个人实现其预期目标的本领，如行动效率。中国古典文献中，孔子“智者若何，仁者若何”（《荀子 • 子道》）的发问也在一定程度上体现了二者的差别；魏征在《谏太宗十思疏》中也论及了这一点：“智者尽其谋……仁者播其惠。”现实生活中，在温暖维度得到较高评价的个体在能力维度未必有同样结果，“老好人”就是一个典型例子，反之亦然。我们认为 Fiske 等（2007）的声誉双维度理论同样适用于第三方惩罚，这引出了本文的研究问题 2：第三方惩罚对惩罚者声誉的两个维度是否具有不同影响？换言之，第三方惩罚是否同时影响了惩罚者声誉的两个维度？影响的方向和程度是否一致？如能回答上述问题，我们就能更为细致地揭示惩罚影响声誉的不同途径。

最后，实验室环境下的第三方惩罚多采用经济惩罚（financial sanction）的形式（陈欣等, 2014），即惩罚者支付一定的金钱成本用以扣减违规者的报酬（Balliet et al., 2011），尽管在不同的研究中金钱成本的支付往往采取不同形式（陈思静等, 2020），但 Guala（2012）指出，这种形式的惩罚很可能只是实验室环境下的人为设定，现实生活中，人们更倾向于运用社会惩罚（social sanction）来维系规范的运作。社会惩罚亦被称为道德惩罚（崔丽莹等, 2017）、非金钱惩罚（Noussair & Tucker, 2005）或流言（Wu, Balliet et al., 2016），其基本形式为人们通过言语来表达对某种违规行为的道德谴责，而不涉及金钱或物质成本（Nelissen & Mulder, 2013; Noussair & Tucker, 2005）。尽管有学者开始考察经济惩罚和社会惩罚对合作或社会规范的影响，但目前尚无研究检验惩罚形式对惩罚者声誉的影响，而现有的关于惩罚者声誉的文献多基于经济惩罚，得出的结论可能在一定程度上存在片面性。由于经济惩罚和社会惩罚在表现形式（物质扣减 vs.言语谴责）、成本（物质成本 vs.非物质成本）、对个体结果（降低受罚者的物质收益 vs.降低受罚者在群体中的名声）和对群体结果（降低群体的净收益 vs.不影响群体净收益）等方面均存在明显差异（Guala, 2012），我们推测这两种惩罚同样会在惩罚者的声誉中产生不同影响。由此，本文的研究问题 3 为：经济惩罚和社会惩罚是否对惩罚者声誉具有不同影响？

2 实验 1

2.1 被试

我们通过软件 G*Power 3.1 来确定所需样本量：取中等效应量 $f^2 = 0.15$，显著性水平 $\alpha = 0.05$，需要 89 名被试才能达到 95% $(1 - \beta)$的统计检验力，而实际参与实验 1 的被试为 90 名

来自某高校非心理学专业的本科生。被试平均年龄为 20.86 ± 1.27 岁，女性占 61.11%，所有被试从未参加过类似实验。被试的专业分布如下：理工科占 36.67%、社会科学占 33.33%、人文学科占 22.22%、艺术及其他占 7.78%。实验开始前，我们通过指导语和练习题确保被试完全了解了实验规则和专业术语的准确含义（例题见附录，余同），并获得了所有被试的知情同意书。

2.2 设计与变量

实验 1 为被试内设计。自变量是惩罚，操作定义是被试扮演第三方时做出的平均惩罚次数。因变量是惩罚者声誉的两个维度（温暖与能力），通过 6 个题项的 Likert 量表来测量。温暖维度的题项包括我觉得某成员：1）值得信赖；2）受人尊敬；3）很友善；而能力维度的题项包括我觉得某成员：4）能为团体带来更多收益；5）其举动对维护团体利益很有帮助；6）可以起到统率团体的作用。题项 1~3 改编自 Barclay（2006），4~6 改编自 Hardy 和 van Vugt（2006），所有题项均为 7 点计分，1 表示完全不同意，7 表示完全同意。调节变量是对惩罚动机的归因，通过 1 个题项来测量：针对某成员的惩罚情况，我认为他/她的这种表现是出于自我聚焦的——集体聚焦的动机。该题项同样为 7 点计分，1 表示完全自我聚焦（self-focused，即关注个人利益），7 表示完全集体聚焦（group-focused，即关注集体利益）。

2.3 程序

实验 1 由 12 轮带有第三方的独裁者博弈组成，通过 z-Tree 上机实验的方式完成（Fischbacher, 2007）。被试被随机分为 30 组，每组 3 人，被试的真实姓名均被 A、B 和 C 等编号所取代。实验期间，被试位于单独隔间内并且不允许相互交流。实验指导语一律采用中性语言（如扣减）来代替带有感情色彩的语言（如惩罚）。实验开始前，告知被试他/她将与其他 2 名成员分别扮演分配者、接受者和第三方（为避免对被试产生潜在暗示，在实际指导语中，分配者、接受者和第三方分别用角色甲、角色乙和角色丙代替，实验 2 亦如此，不再赘述）。在每一轮博弈中，被试随机扮演分配者、接受者或第三方的角色，但在整个实验中，每个被试扮演每个角色的总次数相等，均为 4 次。每一轮博弈开始时，分配者都从实验者手里获得 10 代币（相当于 30 人民币）的初始金额，而第三方和接受者分别获得 5 和 0 代币。分配者根据自己意愿将随意比例的金额分配给接受者，而接受者无法反对，无论分配方案是否公平。如果第三方认为分配方案不公平，可惩罚分配者，惩罚规则统一为第三方付出 2 代币扣减分配者 6 代币。

实验开始后，分配者对初始金额进行分配，第三方看到分配方案后选择是否进行惩罚，然后，分配方案以及第三方的惩罚决定呈现在每个被试的屏幕上。最后一轮博弈结束后，实

验者向每个被试反馈同组其他 2 名成员在 12 轮博弈中的表现，包括：1）作为第三方时做出的平均惩罚次数；2）作为接受者时接受到的平均金额；3）实验结束时手中的代币总数。12 轮博弈结束后，被试使用前文提及量表逐一评价同组其他 2 名成员，包括温暖、能力和惩罚动机。完成上述步骤后，实验者向被试解释实验目的并支付实验报酬，报酬由出场费和 12 轮实验中随机抽取一轮后被试手中的代币组成。

2.4 结果与讨论

对这两个维度的 6 个题项进行验证性因子分析，预期的二因子模型显示出较高的拟合度（CMIN/DF = 3.020，RMSEA = 0.048，GFI = 0.991， CFI = 0.997，NFI = 0.995，PNFI = 0.531，PGFI = 0.378），且二因子模型显著优于（$\Delta\chi^2/df$ = 366.461，$\boldsymbol{p}$ < 0.001）单因子模型（CMIN/DF = 43.402，RMSEA = 0.219，GFI = 0.846， CFI = 0.924，NFI = 0.922，PNFI = 0.553，PGFI = 0.362）。表 1 描述了各变量的均值、标准差和相关系数。不同性别（$\boldsymbol{F}$ = 0.03~1.28, $\boldsymbol{p}$ = 0.261~0.864）和专业（$\boldsymbol{F}$ = 0.48~1.45, $\boldsymbol{p}$ = 0.197~0.846）下惩罚、归因、温暖和能力四个主要变量的差异不显著。

表 1 变量描述性统计与相关系数

变量	***M***	***SD***	1	2	3	4	5
1 惩罚	0.24	0.29					
2 归因	3.88	1.78	0.06				
3 总钱数	54.80	8.34	0.03	0.04			
4 被分配钱数	3.49	1.34	0.16	0.09	0.71^{**}		
5 能力	3.07	1.86	0.58^{**}	0.48^{**}	0.10	0.15	
6 温暖	3.00	1.42	-0.18	0.53^{**}	0.03	0.12	0.24^{*}

注：N = 90，$^{**}p$ < 0.01，$^{*}p$ < 0.05。

以能力为因变量，采用层次回归对惩罚和归因的主效应及调节效应进行检验，为降低多重共线性，对自变量、调节变量和控制变量均进行了中心化处理。回归分析结果如表 2 所示：在模型 M_1 中惩罚（$\boldsymbol{B}$ = 3.52，β = 0.55，$\boldsymbol{p}$ < 0.001，95%CI = [2.59，4.46]）和归因（$\boldsymbol{B}$ = 0.47，β = 0.45，$\boldsymbol{p}$ < 0.001，95%CI = [0.32，0.62]）对能力的主效应都显著：被试做出的惩罚次数越多或惩罚被归因为集体聚焦的程度越高，获得的能力评价就越高。上述结果表明，惩罚与动机归因均会显著影响他人对惩罚者能力的评价。

表 2 层次回归对主效应和调节效应的检验（能力维度）

	M_1				M_2				M_3			
	B	*SE*	***t***	***p***	***B***	*SE*	***t***	***p***	***B***	*SE*	***t***	***p***
主效应												
惩罚	3.52	0.47	7.50	0.000	3.44	0.41	8.43	0.000	3.56	0.41	8.64	0.000
归因	0.47	0.08	6.09	0.000	0.38	0.07	5.46	0.000	0.38	0.07	5.52	0.000
调节效应												
惩罚×归因					1.19	0.22	5.41	0.000	1.26	0.22	5.63	0.000
控制变量												
被分配钱数									-0.22	0.13	-1.70	0.093
总钱数									0.03	0.02	1.60	0.113
R^2	0.53^{***}				0.65^{***}				0.66^{***}			
ΔR^2					0.12^{***}				0.01			

注：$N = 90$，$^{***}p < 0.001$，$^{*}p < 0.05$。

在模型 M_2 中惩罚与归因的交互项对能力有显著的正向影响（$\boldsymbol{B} = 1.19$，$\boldsymbol{\beta} = 0.36$，$\boldsymbol{p} < 0.001$，95%CI = [0.75，1.62]），可以解释能力变异量的 12%。这说明惩罚对能力的影响受到归因的正向调节作用。为了更清晰地显示归因的调节作用，用 Johnson-Neyman 法进一步量化分析归因对惩罚与能力关系的影响，并检验调节效应的统计显著区，结果如图 1 所示。

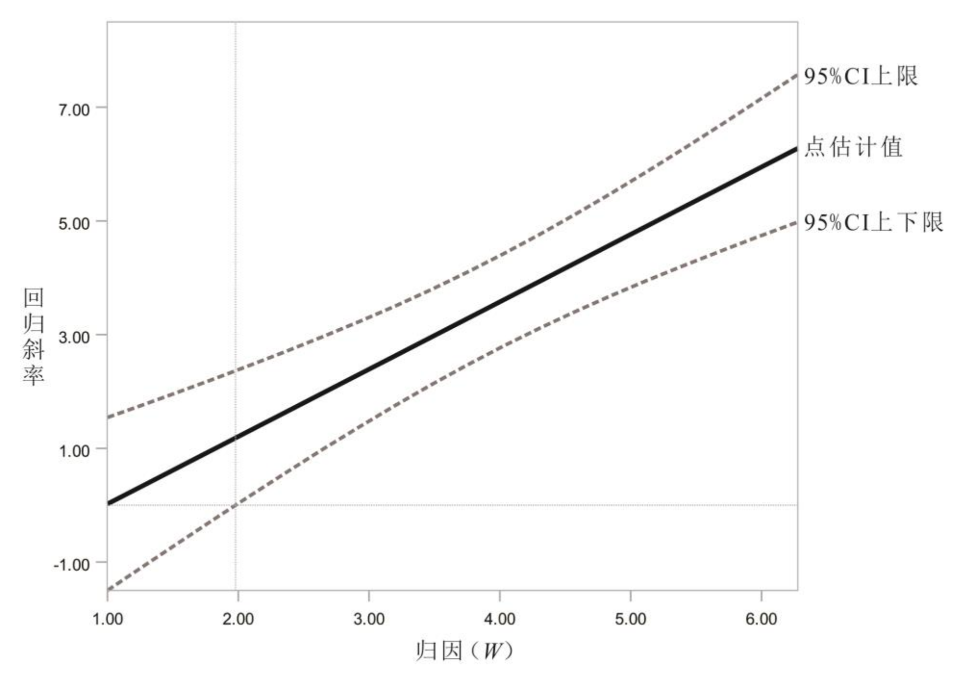


图 1 归因对惩罚和能力之间关系的影响

通过图 1 可以看出，当归因超过 2 时，惩罚影响能力的回归斜率置信区间都在 0 点以上，说明当惩罚动机归因超过上述阈值时，被归因为集体聚焦程度越高的惩罚对能力的提升作用也就越大；而当归因低于 2 时，置信区间包含 0 点，此时惩罚对能力的影响不显著。上述结果表明惩罚对能力的影响是有条件的，归因为自我聚焦的惩罚易被感知为一种自利手段，而非维护社会规范的行为，因而不大可能对集体利益产生积极的影响，从而失去了提升能力评价的作用。因此有理由认为，被看作聚焦于集体利益的惩罚才可能提升惩罚者的能力评价。

以声誉的另一个维度——温暖为因变量，采用相同方法对惩罚和归因的主效应及调节效应进行检验，结果如表 3 所示：在模型 M_1 中惩罚（$\boldsymbol{B}$ = -1.24，$\boldsymbol{\beta}$ = -0.27，$\boldsymbol{p}$ = 0.003，95%CI = [-2.05，-0.44]）和归因（$\boldsymbol{B}$ = 0.42，$\boldsymbol{\beta}$ = 0.55，$\boldsymbol{p}$ < 0.001，95%CI = [0.29，0.55]）对温暖的主效应都显著：被试做出的惩罚次数越多，获得的温暖评价就越低；惩罚被归因为集体聚焦的程度越高，获得的温暖评价就越高。上述结果表明，惩罚会显著降低对温暖的评价，而偏于集体聚焦的归因有助于减缓这种负面影响。

表 3 层次回归对主效应和调节效应的检验（温暖维度）

	M_1				M_2				M_3			
	B	*SE*	***t***	***p***	***B***	*SE*	***t***	***p***	***B***	*SE*	***t***	***p***
主效应												
惩罚	-1.24	0.40	-3.09	0.003	-1.28	0.39	-3.27	0.002	-1.39	0.40	-3.48	0.001
归因	0.42	0.07	6.32	0.000	0.38	0.07	5.69	0.000	0.38	0.07	5.66	0.000
调节效应												
惩罚×归因					0.52	0.21	2.48	0.015	0.45	0.22	2.09	0.039
控制变量												
被分配钱数									0.18	0.13	1.43	0.156
总钱数									-0.02	0.02	-1.04	0.300
R^2	0.59^{***}				0.63^{***}				0.64^{***}			
ΔR^2					0.04^{*}				0.01			

注：$N = 90$，$^{***}p < 0.001$，$^{*}p < 0.05$。

在模型 M_2 中惩罚与归因的交互项对温暖有显著的正向影响（$\boldsymbol{B}$ = 0.52，$\boldsymbol{\beta}$ = 0.22，$\boldsymbol{p}$ = 0.015，95%CI = [0.10，0.94]），可以解释温暖变异量的 4%。这说明惩罚对温暖的影响受到

归因的调节作用。用 Johnson-Neyman 法进一步量化分析归因对惩罚与温暖关系的影响，并检验调节效应的统计显著区，结果如图 2 所示。

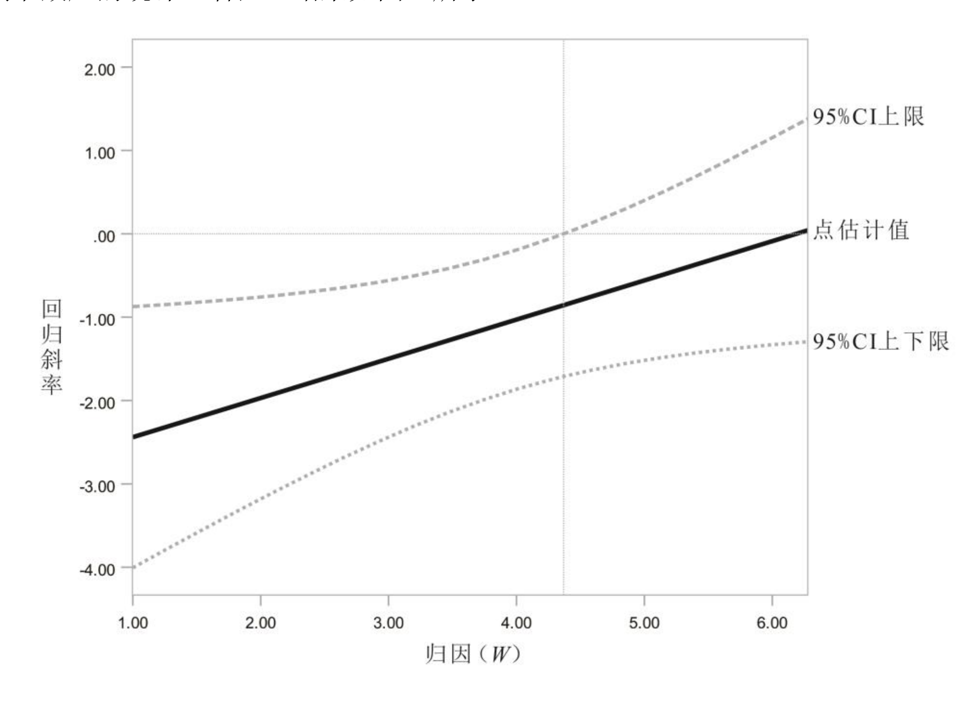


图 2 归因对惩罚和温暖之间关系的影响

通过图 2 可以看出，当归因低于 4.39 时，惩罚影响温暖的回归斜率置信区间都在 0 点以下，这说明当归因低于上述阈值时，惩罚被归因为个体聚焦的程度越高，惩罚降低温暖评价的作用也就越大；而当归因高于 4.39 时，置信区间包含 0 点，此时惩罚对温暖的影响不显著，这说明当动机归因偏向集体聚焦时，惩罚降低温暖的负面作用就消失了。上述结果表明，惩罚大体上会降低我们对惩罚者在温暖维度上的评价，但归因为集体聚焦的惩罚会被感知为一种维护群体规范、提升群体利益的行为，从而消除了对温暖的负面作用。因此有理由认为，只要惩罚在足够高的程度上被认为是出于维护集体利益，惩罚就不会降低惩罚者的温暖评价。

实验 1 的结果为研究问题 1 和 2 提供了初步回答。从实验 1 的结果中我们可以得出两个重要结论：1）惩罚行为对惩罚者声誉的两个维度具有显著不同的作用，简言之，惩罚在总体上提高惩罚者的能力评价而降低其温暖评价。这意味着惩罚对声誉两个维度的影响方向是相反的，这在一定程度上解释了先前研究中看似矛盾的发现，如 de Kwaadsteniet 等（2019）注意到，比起从不惩罚失职员工的领导来，人们对采取惩罚措施的领导的评价更高，但矛盾的是，人们却更喜欢前者；Barclay（2006）也有类似的发现。基于实验 1 的结果，我们认为这是因为惩罚提高惩罚者能力评价的同时也降低了其温暖评价。2）对惩罚动机的归因显

著影响了他人对惩罚者声誉的评价，具体而言，惩罚动机越表现为关注集体利益，就越能提升惩罚对能力的正面影响，同时也越能降低对温暖的负面影响。这也意味着惩罚与社会规范之间可能存在双向的作用机制，尽管目前相关研究多关注第三方惩罚对社会规范的维系作用（Fehr & Fischbacher, 2004; Fehr & Gächter, 2002），但也有学者指出，当缺乏社会规范的合理指引时，惩罚会对合作产生负面作用（Bicchieri et al., 2018; Fehr & Rockenbach, 2003），实验 1 表明上述效应同样存在于惩罚者的声誉里，即只有被认为是符合规范的惩罚（关注集体利益）才有可能在总体上提升惩罚者的声誉。

3 实验 2

实验 1 表明第三方惩罚对惩罚者声誉的两个维度具有不同影响，同时人们对惩罚动机的归因会影响对惩罚者声誉的评价。尚需进一步回答的问题是人们依据什么对惩罚动机进行归因。正如 Kiyonari 和 Barclay（2008）指出，现实生活中旁观者不大可能完全了解惩罚的前因后果，旁观者需要通过可用线索来判断惩罚动机。在实验 2 中，我们引入了被试可据以判断惩罚动机的信息线索。我们推测，惩罚者本人合作水平的高低在一定程度上暗示了其惩罚动机是否合理，例如，对公共物品从无贡献或者在分配资源时表现吝啬的个体似乎很难认为其惩罚动机是为了维护某种规范。此外，尽管经济惩罚是目前第三方惩罚实验室研究的主流，但正如 Guala（2012）注意到，和实验室环境相反，现实生活中人们更愿意采用社会惩罚而不是经济惩罚去惩戒违规者。因此，实验 2 的另一个目的是引入经济惩罚和社会惩罚两种惩罚形式并考察它们对惩罚者声誉的影响。

3.1 被试

实验 2 使用二元三因素方差分析检验自变量的主效应和交互作用。我们通过软件 G*Power 3.1 确定样本量：取中等效应量 $f^2 = 0.0625$，显著性水平 $\alpha = 0.05$，需要 171 名被试才能达到 95% (1 − β)的统计检验力，而实际共有 176 名社会被试参与了实验 2。被试平均年龄为 35.07 ± 17.49 岁，女性占 59.66%；职业分布为：学生占 25.57%，机关及事业单位占 18.75%，各类企业占 24.43%，个体经营占 19.32%，其它占 11.93%；受教育程度分布为：中专及以下占 27.27%，大专占 21.59%，本科占 45.45%，硕士和博士占 5.68%；月收入分布为：2000 元以下占 10.80%，2000~5000 元占 28.41%，5000~1 万元占 44.89%，1 万元以上占 15.91%。所有被试之前均未参加过类似实验并在实验开始前均已签署知情同意书。

3.2 设计与变量

实验 2 是 2（合作：低/高）×2（经济惩罚：无/有）×2（社会惩罚：无/有）的被试内设计。合作的操作定义是被试作为分配者时分配给接受者的金额；经济惩罚的操作定义是被试付出 2 代币扣减分配者 6 代币；社会惩罚的操作定义是被试向分配者发送信息，信息为“我认为你的分配方案不公平”（Nelissen & Mulder, 2013）。和实验 1 一样，因变量是惩罚者声誉的两个维度，通过 6 个题项的 Likert 量表来测量（见实验 1）。

3.3 程序

实验 2 仍然是 12 轮带有第三方的独裁者博弈，程序和实验 1 大致相似，除了：1）告知被试他/她将与其他 8 名成员分别扮演分配者、接受者和第三方，但事实上，其他 8 名成员并非真实被试，而是实验者事先设定的程序；2）告知被试，每一轮博弈中每个小组 9 名成员将随机分成 3 个分组，每个分组中都有 1 名分配者、1 名接受者和 1 名第三方，并且，当被试扮演第三方时，实验者向其反馈所在分组的分配方案；而当被试扮演其他角色时，本轮无信息反馈，这样安排的目的是尽管每一轮博弈都是以 3 名成员为单位展开，但由于每一轮博弈中 3 个分组都是随机组成的，因此被试有同等几率与同组其他 8 名虚拟成员进行直接互动；3）面对第三方认为不公平的分配方案，第三方可选择不惩罚、经济惩罚、社会惩罚或同时实施两种惩罚；4）在每一轮博弈中，被试随机扮演分配者、接受者或第三方的角色，但在整个实验中，每个被试扮演每个角色的总次数相等，均为 4 次；5）最后一轮博弈结束后，实验者向每个被试反馈同组其他 8 名成员在 12 轮博弈中的表现，包括：①作为分配者分配给接受者的金额水平（低/高）；②作为第三方是否做出过经济惩罚（无/有）；③作为第三方是否做出过社会惩罚（无/有）。事实上，反馈由实验者事先设定，包含 2（合作：低/高）×2（经济惩罚：无/有）×2（社会惩罚：无/有）这 8 种情况，每种情况对应 1 名成员。所有被试看到的反馈都是相同的，但按随机顺序呈现。接着，被试评价同组其他 8 名成员，并对这些成员的惩罚情况进行归因。评价和归因所使用量表与实验 1 相同。完成上述步骤后，实验者宣布实验结束，并向被试解释实验目的和支付实验报酬。

3.4 结果与讨论

首先检验合作高低是否显著影响了被试对惩罚的归因：被试对高合作者惩罚的归因（M = 3.01，SD = 1.45）显著高于低合作者（M = 2.45，SD = 1.81）（t = 6.46，p < 0.001，d = 0.34，95%CI = [0.24，0.45]），这表明高合作者的惩罚更可能被归因为集体聚焦，因此和我们预测的一样，惩罚者的合作行为确实是一种重要的归因线索。不同性别、职业、受教育程度和收入水平下温暖（F = 0.23~1.01, p = 0.463~0.921）和能力（F = 0.

60~1.64, p = 0.07~0.62）两个主要变量的差异不显著，年龄与温暖（r = -0.04, p = 0.635）和能力（r = 0.09, p = 0.247）的相关系数都不显著。表 4 展示了描述统计结果。

表 4 温暖与能力的描述统计结果

合作	社会惩罚	经济惩罚	温暖	能力
低	无	无（N = 22）	3.60（1.55）	3.17（1.60）
		有（N = 22）	3.36（1.38）	3.50（1.44）
	有	无（N = 22）	3.37（1.44）	3.44（1.52）
		有（N = 22）	3.06（1.66）	3.24（1.75）
高	无	无（N = 22）	4.89（1.26）	4.06（1.60）
		有（N = 22）	4.43（1.26）	4.30（1.44）
	有	无（N = 22）	4.92（1.22）	4.85（1.23）
		有（N = 22）	4.47（1.37）	4.55（1.52）

注：表中数值为平均数与标准差。

以温暖和能力为因变量，合作、社会惩罚和经济惩罚为自变量进行二元三因素方差分析。多变量检验结果显示，合作（Wilks’ Lambda = 0.82，F = 157.17，p < 0.001，偏 η^2 = 0.18）、社会惩罚（Wilks’ Lambda = 0.97，F = 22.77，p < 0.001，偏 η^2 = 0.03）和经济惩罚（Wilks’ Lambda = 0.96，F = 29.04，p < 0.001，偏 η^2 = 0.04）对两个因变量的主效应显著；合作与社会惩罚（Wilks’ Lambda = 0.99，F = 5.15，p = 0.006，偏 η^2 = 0.01）以及经济惩罚与社会惩罚（Wilks’ Lambda = 0.99，F = 10.99，p < 0.001，偏 η^2 = 0.02）的交互作用也显著；而合作与经济惩罚（Wilks’ Lambda = 0.99，F = 0.88，p = 0.415）及三者（Wilks’ Lambda = 1，F = 0.08，p = 0.929）的交互作用不显著。这说明总体上社会惩罚和经济惩罚都会直接影响声誉，同时，在不同合作水平下社会惩罚对声誉的影响有所不同，而在不同社会惩罚水平下经济惩罚对声誉的影响也不同。

进一步对主体间效应进行检验，结果如表 5 所示。分析结果显示，合作对温暖和能力的主效应都显著；经济惩罚对能力的主效应不显著，对温暖的主效应显著；社会惩罚对能力的主效应显著，对温暖的主效应不显著；此外，合作与社会惩罚的交互作用在能力和温暖两个维度上都显著，经济惩罚与社会惩罚的交互作用在能力维度上显著。

表 5 二元三因素方差分析结果

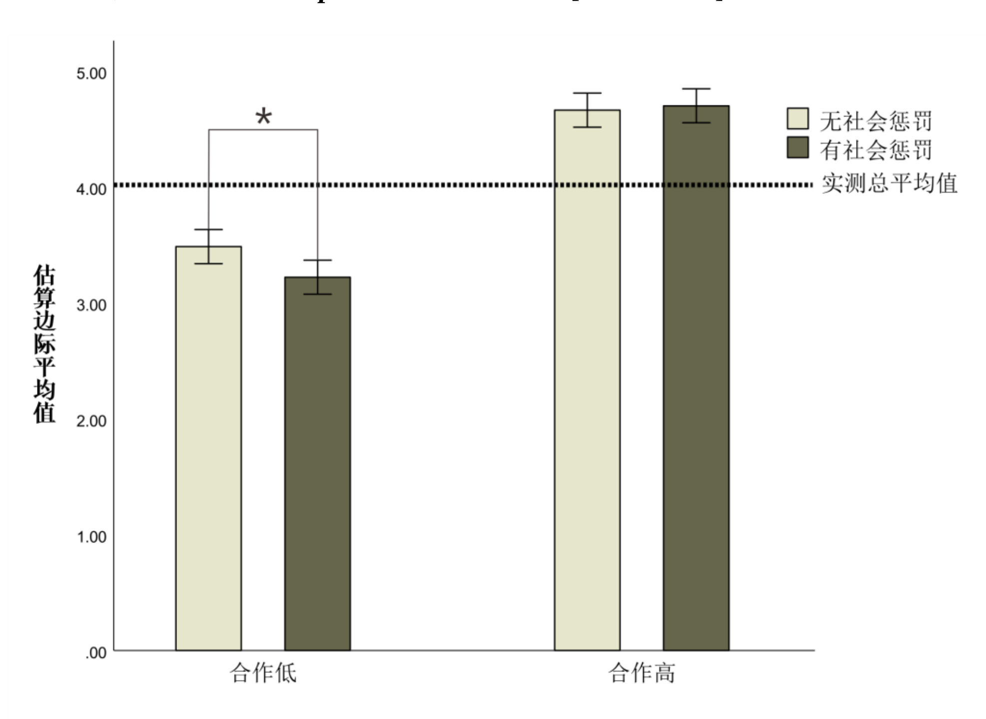
来源	因变量	均方	F	显著性	偏 η^2

修正模型	温暖 [a]	97.79	49.50	0.000	0.20
	能力 [b]	71.85	30.96	0.000	0.13
截距	温暖	22728.44	11504.62	0.000	0.89
	能力	21341.39	9196.93	0.000	0.87
合作	温暖	620.90	314.29	0.000	0.18
	能力	429.52	185.10	0.000	0.12
经济惩罚	温暖	47.89	24.24	0.000	0.02
	能力	0.08	0.03	0.857	0.00
社会惩罚	温暖	4.51	2.28	0.131	0.00
	能力	24.31	10.48	0.001	0.01
合作×经济惩罚	温暖	2.94	1.49	0.223	0.00
	能力	0.74	0.32	0.572	0.00
合作×社会惩罚	温暖	8.03	4.07	0.044	0.003
	能力	23.27	10.03	0.002	0.01
经济惩罚×社会惩罚	温暖	0.11	0.06	0.815	0.00
	能力	25.01	10.78	0.001	0.01
合作×社会惩罚×经济惩罚	温暖	0.13	0.07	0.796	0.00
	能力	0.00	0.00	0.995	0.00

a: R^2 = 0.198（调整后 R^2 = 0.194） b: R^2 = 0.134（调整后 R^2 = 0.130）

合作与社会惩罚的交互作用在能力和温暖两个维度上显著，因此进一步分析在不同合作水平下社会惩罚的简单效应。多变量检验结果显示，低合作水平（Wilks’ Lambda = 0.99，F = 7.27，p = 0.001，偏 η^2 = 0.01）和高合作水平（Wilks’ Lambda = 0.97，F = 20.65，p < 0.001，偏 η^2 = 0.03）下，社会惩罚对两个因变量的简单效应都显著，从效应量上来看，在高合作水平下，社会惩罚对声誉的影响更大。对温暖和能力两个维度做单变量检验结果显示：低合作水平下社会惩罚对温暖的简单效应显著（F = 6.22，p = 0.013，偏 η^2 = 0.004），高合作水平下社会惩罚对温暖的简单效应不显著（F = 0.13，p = 0.721），说明低合作成员做出的社会惩罚会显著降低温暖评价，而高合作成员做出的社会惩罚不会对温暖有负面影响；低合作水平下社会惩罚对能力的简单效应不显著（F = 0.002，p = 0.961），高合作水平下社会惩罚对能力的简单效应显著（F = 20.50，p < 0.001，偏 η^2 = 0.01），说明只有高合作

成员做出的社会惩罚才能显著提升能力评价。成对比较（Bonferroni 法校正）的结果进一步验证了上述判断（见图 3 和图 4）：在低合作水平下，相比于不做社会惩罚（***M*** = 3.48，***SE*** = 0.08）的成员，被试对做出社会惩罚（***M*** = 3.22，***SE*** = 0.08）的成员的温暖评价显著偏低（***p*** = 0.013，95%CI = [0.06，0.47]）；在高合作水平下，与不做出社会惩罚（***M*** = 4.66，***SE*** = 0.075）的成员相比，被试对做出社会惩罚（***M*** = 4.70，***SE*** = 0.08）的成员温暖评价并未显著降低（***p*** = 0.721，95%CI = [-0.25，0.17]）。在高合作水平下，相比于不做社会惩罚（***M*** = 4.18，***SE*** = 0.08）的成员，被试对做出社会惩罚（***M*** = 4.70，***SE*** = 0.08）的成员能力评价显著提高（***p*** < 0.001，95%CI = [-0.75，-0.30]）；在低合作水平下，与不做出社会惩罚（***M*** = 3.33，***SE*** = 0.08）的成员相比，被试对做出社会惩罚（***M*** = 3.34，***SE*** = 0.08）的成员能力评价并未显著提高（***p*** = 0.961，95% CI = [-0.23，0.22]）。



注：$^{***}p < 0.001$，$^{**}p < 0.01$，$^{*}p < 0.05$，下同

图 3 不同合作水平下社会惩罚对温暖的影响

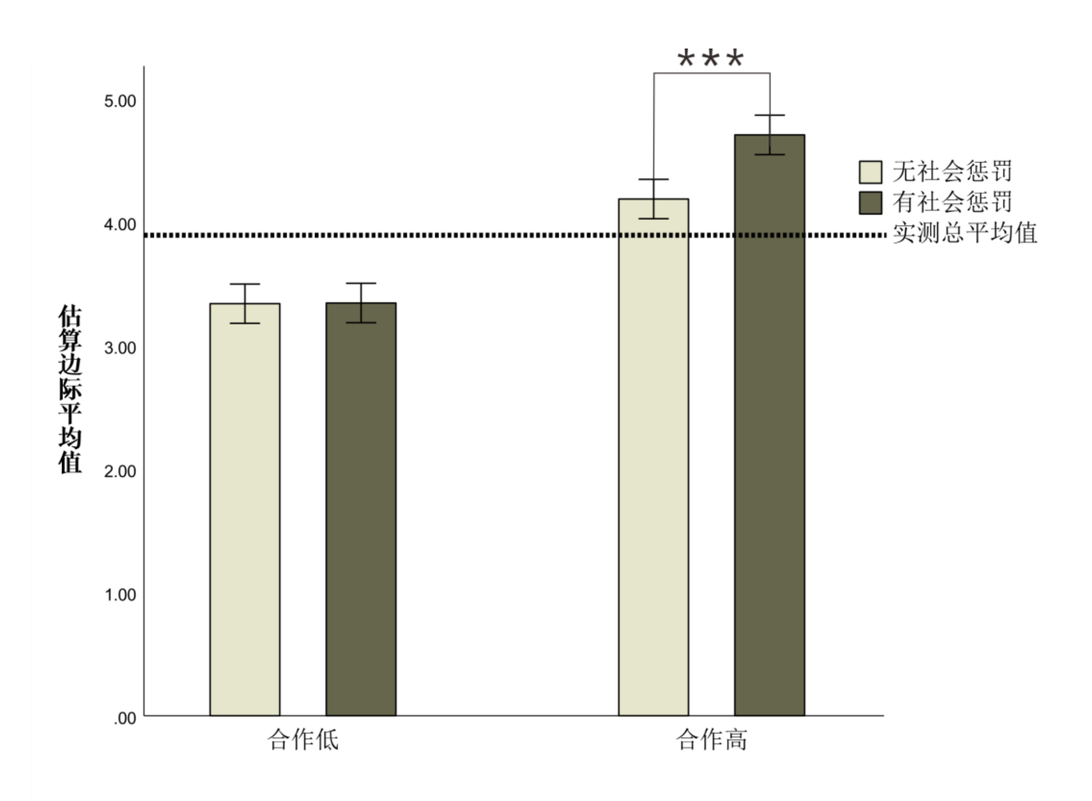


图 4 不同合作水平下社会惩罚对能力的影响

另一方面，社会惩罚与经济惩罚的交互作用在能力维度上显著，因此需进一步分析在不同社会惩罚水平下经济惩罚的简单效应。多变量检验结果显示，无社会惩罚（Wilks' Lambda = 0.96，F = 33.06，p < 0.001，偏 η^2 = 0.05）和有社会惩罚时（Wilks' Lambda = 0.99，F = 6.97，p = 0.001，偏 η^2 = 0.01），经济惩罚对两个因变量的简单效应都显著，从效应量上来看，在无社会惩罚时，经济惩罚对声誉的影响更大。对温暖和能力两个维度做单变量检验结果显示：无社会惩罚（F = 11.00，p = 0.001，偏 η^2 = 0.01）和有社会惩罚时（F = 13.30，p < 0.001，偏 η^2 = 0.01）经济惩罚对温暖的简单效应均显著，说明不管有没有做出社会惩罚，经济惩罚均会显著降低对温暖的评价；无社会惩罚（F = 6.00，p = 0.014，偏 η^2 = 0.004）和有社会惩罚时（F = 4.81，p = 0.028，偏 η^2 = 0.003）经济惩罚对能力的简单效应均显著，说明不管有没有做出社会惩罚，经济惩罚均会影响能力评价，但影响的方向不同。成对比较（Bonferroni 法校正）的结果进一步表明（见图 5 和图 6）：在无社会惩罚时，相比于不做经济惩罚（M = 4.25，SE = 0.08）的成员，被试对做出经济惩罚（M = 3.89，SE = 0.08）的成员温暖评价显著降低（p = 0.001，95%CI = [0.14，0.56]）；在有社会惩罚时，与不做经济惩罚（M = 4.15，SE = 0.08）的成员相比，被试对做出经济惩罚（M = 3.76，SE = 0.08）的成员温暖评价也显著降低（p < 0.001，95%CI = [0.18，0.59]）。在无社会惩罚时，相比于不做经济惩罚（M = 3.62，SE = 0.08）的成员，被试对做出经济惩罚（M = 3.90，SE = 0.081）

的成员能力评价显著提高（***p*** = 0.014，95%CI = [-0.51，-0.06]）；在有社会惩罚时，与不做经济惩罚（***M*** = 4.15，***SE*** = 0.08）的成员相比，被试对做出经济惩罚（***M*** = 3.89，***SE*** = 0.08）的成员能力评价显著降低（***p*** = 0.028，95%CI = [0.03，0.48]）。也就是说，在无/有社会惩罚时，经济惩罚对能力维度的作用方向是相反的。

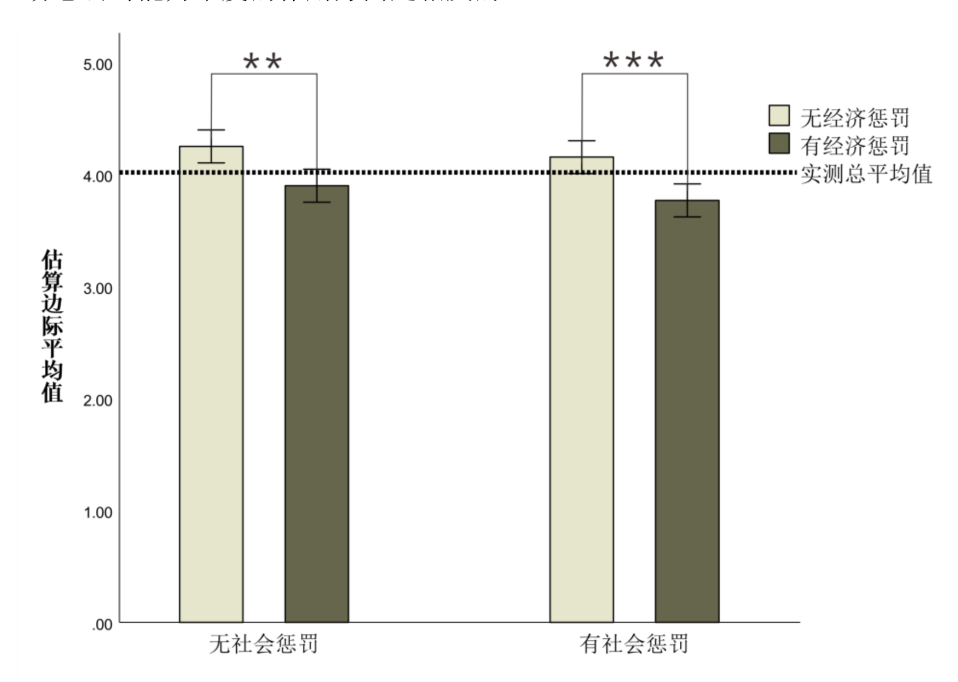


图 5 无/有社会惩罚时经济惩罚对温暖的影响

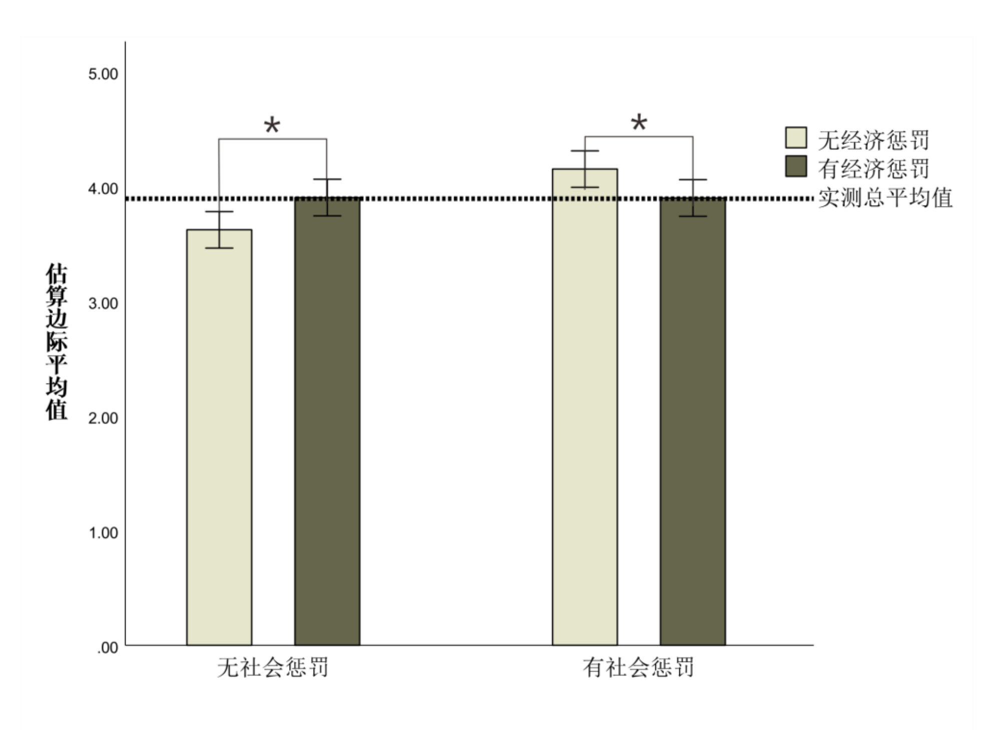


图 6 无/有社会惩罚时经济惩罚对能力的影响

实验 1 初步回答了本文的研究问题 1，即旁观者对惩罚动机的归因是否对声誉的两个维度具有不同影响，实验 2 一方面进一步拓展了研究问题 1，即旁观者依据什么线索来对惩罚动机进行归因。结果发现，惩罚者本人的合作水平是他人进行动机归因的一种重要线索：惩罚者的高合作水平使得旁观者倾向于将其惩罚动机归因为关注集体利益，而低合作水平则是惩罚者关注私利的一个信号。实验 1 表明，惩罚总体上降低惩罚者的温暖评价而提高其能力评价，被归因为集体聚焦的惩罚能显著降低对温暖的负面作用并进一步提高对能力的正面作用。在实验 2 中，我们也观察到了类似的结果：高合作惩罚者做出的社会惩罚并不影响其温暖评价，但提升了其能力评价。

另一方面，实验 2 的结果也部分地回答了研究问题 3：不同形式的惩罚对声誉是否具有不同影响？实验 2 的结果表明回答是肯定的：经济惩罚不显著影响能力评价，但降低温暖评价；而社会惩罚不显著影响温暖评价，但提升能力评价。此外，对社会惩罚和经济惩罚的交互分析支持了现有文献的一个结论，即经济惩罚经常产生副作用（陈思静，朱玥，2020; Houser et al., 2008），尤其在个体具备其他选项时（谢东杰，苏彦捷, 2019）。在实验 2 中，经济惩罚的副作用主要表现为无论个体是否做出社会惩罚，经济惩罚都降低了温暖评价，这可能是因为社会惩罚被认为是维护社会规范的更好选项（崔丽莹等, 2017）。在存在社会惩罚这个选项的情况下，被试仍然采用经济惩罚可能会被认为是出于负面动机，如自利与恶意（Fehr & Rockenbach, 2003），从而降低了惩罚者的温暖评价。能力维度上的情况有所不同，在社会惩罚缺席的情况下，经济惩罚提高了能力评价，反之，则降低了能力评价。我们推测这和惩罚的效果与效率有关（Balliet et al., 2011）：惩罚效果指惩罚是否能提高合作水平，而惩罚效率是指扣除惩罚成本后惩罚是否能提高集体的净收益。在缺乏社会惩罚的情况下，经济惩罚客观上起到了向违规者提示社会规范的作用（无论其惩罚动机如何），在一定程度上促进了违规者（未来的）合作行为（Bicchieri et al., 2018; Chen et al., 2020），即从效果的角度来讲，经济惩罚客观上具有正面作用，因此提高了惩罚者的能力评价；相反，在社会惩罚已经起到了提示社会规范的情况下，经济惩罚这方面的额外作用可能未必显著，同时，由于经济惩罚的成本较高，最终反而降低了集体的净收益（Dreber et al., 2008），即从效率的角度来说，经济惩罚此时具有潜在的负面作用，从而降低了惩罚者的能力评价。

4 实验 3

实验 1 和 2 基本回答了本文提出的三个研究问题，然而，为了进一步考察本文主要变量之间的关系，我们尚需检验对惩罚动机的归因与惩罚形式间的交互机制，实验 3 旨在解决这一问题。此外，前两个实验已经检验了惩罚者在扮演其他角色时的表现对其声誉的影响（实验 1 显示被试扮演接受者时的表现并不影响其作为惩罚者的声誉，而实验 2 表明被试扮演分配者时的表现显著影响了其作为惩罚者的声誉），在实验 3 中，我们对经典的独裁者博弈范式进行了适当修改，我们不再检验惩罚者的角色效应，而是集中探讨惩罚形式与归因之间的交互作用。最后，由于前两个实验已经考察了是否采取惩罚对声誉的影响，实验 3 不再设置“不惩罚”这一选项。

4.1 被试

我们通过软件 G*Power 3.1 来确定样本量：取中等效应量 $\boldsymbol{f}^2 = 0.15$，显著性水平 $\boldsymbol{\alpha} = 0.05$，需要 119 名被试才能达到 95% (1 − β)的统计检验力，而实际共有 120 名来自某高校非心理学专业本科生参加了实验 3。被试平均年龄为 21.20 ± 1.72 岁，女性占 53.33%；专业分布为：理工科 32.50%，社科类 29.17%，人文学科 20.83%，艺术类及其它 17.50%。

4.2 设计与变量

实验 3 为被试内设计。自变量是惩罚形式（经济惩罚 vs. 社会惩罚，操作定义同实验 2）。因变量是惩罚者声誉的两个维度，调节变量是对惩罚动机的归因。测量声誉的量表同实验 1；测量归因所使用的问题为：针对某成员的惩罚情况，我认为他/她的这种表现是出于关心自我利益（自我聚焦）——关心集体利益（集体聚焦）。该问题为 7 点计分，1 表示完全自我聚焦，7 表示完全集体聚焦。

4.3 程序

实验 3 采用了带有多名第三方的独裁者博弈（Ouss & Peysakhovich, 2015），即在每个小组除了 1 名分配者（角色 A）和 1 名接受者（角色 B）外，有 2 名第三方（角色 C 和 D）。博弈开始前，分配者、接受者和 2 名第三方分别拥有 10、0 和 5 代币的初始金额，分配者在自己与接受者之间自由分配这笔初始金额，接受者无权干预，但 2 名第三方均可对其认为的不公平分配进行惩罚，惩罚分经济惩罚和社会惩罚两个水平。实验 3 中所有被试均为旁观者，不直接参与博弈，而是旁观 1 轮上述由 4 名个体参与的独裁者博弈，他们的任务是在上述博弈完成后，尽快计算出每个个体的最终收益。直接参与博弈的 4 名个体实际上是实验者事前设定的程序。博弈开始后，所有被试看到的分配方案均为分配者将 2 代币（即初始金额的 20%）分配给了接受者，并且第三方 C 和第三方 D 分别对分配者进行了经济惩罚和社会惩

罚。接着，被试计算每个个体的收益，对分配者和 2 名第三方[1]进行评价，并对 2 名第三方的惩罚动机进行归因。

4.4 结果与讨论

在实验 3 中，不同性别（F = 1.09~1.47, p = 0.23~0.30）和专业（F = 1.38~1.42, p = 0.238~0.740）下归因、温暖和能力三个主要变量的差异并不显著。三个变量的描述统计如表 6 所示。

表 6 变量描述性统计与相关系数

变量	***M***	***SD***	1	2	3
1 归因	3.67	1.61			
2 温暖	4.46	1.43	0.16*		
3 能力	3.87	1.63	0.27**	0.36**	

注：N = 90，$^{**}p < 0.01$，$^{*}p < 0.05$。

以惩罚形式（经济惩罚=1，社会惩罚=0）为自变量、归因为调节变量、温暖为因变量做分层回归检验归因的调节作用，结果如表 7 所示。模型 M_1 中惩罚形式（B = 0.80，β = 0.28，p < 0.001，95%CI = [0.45，1.14]）和归因（B = 0.12，β = 0.14，p = 0.027，95%CI = [0.01，0.23]）的主效应都显著，这说明惩罚形式和归因都能显著影响被试对惩罚者温暖的评价。模型 M_2 惩罚形式与归因的交互作用也显著（B = -0.30，β = -0.47，p = 0.005，95%CI = [-0.52，-0.09]），加入交互项后 R^2 变化量显著增加，可以解释温暖变异量的 3%。交互项的系数为负，这说明随着被试将惩罚归因为集体聚焦的倾向越来越强，惩罚形式对温暖的影响逐渐变小。为了更清晰地显示归因的调节作用，用 Johnson-Neyman 法进一步量化分析归因对惩罚形式与温暖关系的影响，并检验调节效应的统计显著区，结果如图 7 所示。

表 7 层次回归对主效应和调节效应的检验（温暖维度）

	M_1				M_2				M_3			
	B	*SE*	***t***	***p***	***B***	*SE*	***t***	***p***	***B***	*SE*	***t***	***p***
主效应												
惩罚形式	0.80	0.18	4.54	0.000	1.92	0.43	4.42	0.000	1.85	0.44	4.25	0.000
归因	0.12	0.06	2.23	0.027	0.27	0.07	3.58	0.000	0.27	0.07	3.60	0.000

[1] 实验 3 的主要目的在于考察旁观者对惩罚者声誉的评价，但为了避免问题显得过于具有诱导性，我们也要求被试评价分配者，但在后续统计分析中，我们主要关注被试对惩罚者的评价。

调节效应												
惩罚形式×归因					-0.30	0.11	-2.81	0.005	-0.28	0.11	-2.59	0.010
控制变量												
性别									-0.20	0.17	-1.15	0.253
专业									0.09	0.08	1.15	0.252
R^2	0.10^{***}				0.13^{***}				0.14^{***}			
ΔR^2					0.03^{**}				0.01			

注：$N = 120$，$^{***}p < 0.001$，$^{**}p < 0.01$。

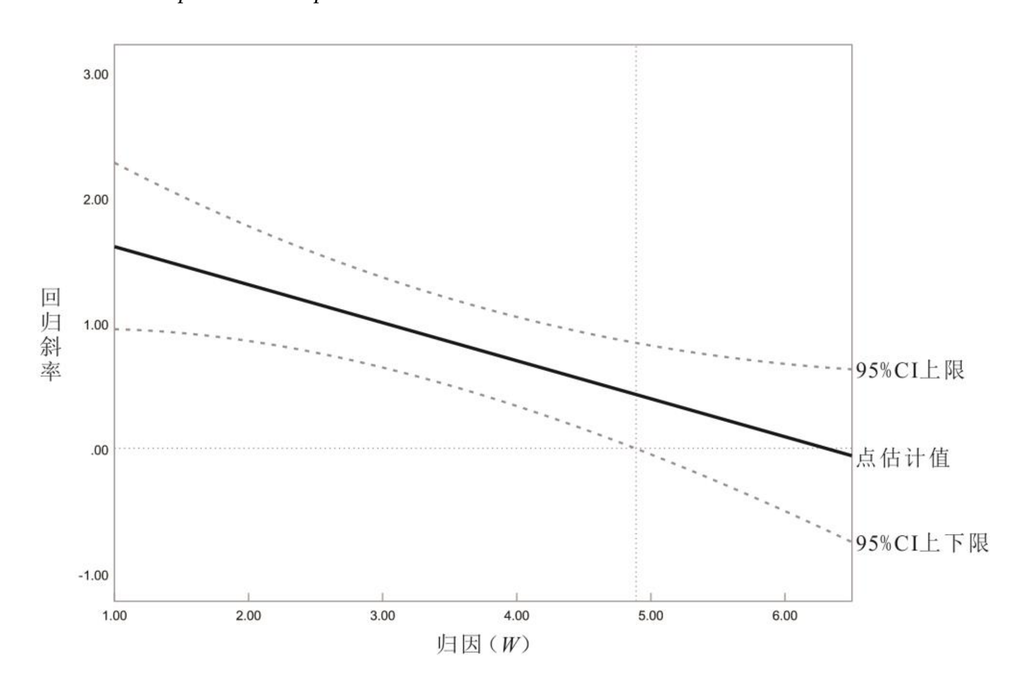


图 7 归因对惩罚形式和温暖关系的调节作用

通过图 7 可以看出，当归因低于 4.89 时，惩罚形式影响温暖的回归斜率置信区间都在 0 点以上，这说明惩罚形式对温暖有显著影响，相对于经济惩罚，被试对做出社会惩罚个体的温暖评价更高；而当归因大于 4.89 时，置信区间包含 0 点，此时惩罚形式对温暖的影响不显著，被试对做出两种惩罚个体的温暖评价无显著差别。上述结果表明惩罚形式对温暖的影响是有条件的：当被试将惩罚归因为自我聚焦时，会对做出经济惩罚的个体更低的温暖评价；当被试将惩罚归因为集体聚焦时，对两种惩罚者的温暖评价无显著差异。

进一步以惩罚形式（经济惩罚=1，社会惩罚=0）为自变量、归因为调节变量、能力为因变量做分层回归检验归因的调节作用，结果如表 8 所示。模型 M_1 中惩罚形式（$\boldsymbol{B} = 0.70$，$\boldsymbol{\beta} = 0.22$，$\boldsymbol{p} < 0.001$，95%CI = [0.31，1.10]）和归因（$\boldsymbol{B} = 0.26$，$\boldsymbol{\beta} = 0.25$，$\boldsymbol{p} < 0.001$，95%CI = [0.13，0.38]）的主效应都显著，这说明惩罚形式和归因都能显著影响被试对惩罚者能力的

评价。模型 M_2 惩罚形式与归因的交互作用也显著（***B*** = 0.51，**β** = 0.69，***p*** < 0.001，95%CI = [0.28，0.75]），加入交互项后 R^2 变化量显著增加，可以解释能力变异量的 6%。交互项的系数为正，这表明随着被试将惩罚归因为集体聚焦的倾向越来越强，惩罚形式对能力的影响逐渐变大。为了更清晰地显示归因的调节作用，用 Johnson-Neyman 法进一步量化分析归因对惩罚形式与能力关系的影响，并检验调节效应的统计显著区，结果如图 8 所示。

表 8 层次回归对主效应和调节效应的检验（能力维度）

	M_1				M_2				M_3			
	B	*SE*	***t***	***p***	***B***	*SE*	***t***	***p***	***B***	*SE*	***t***	***p***
主效应												
惩罚形式	0.70	0.20	3.54	0.000	1.18	0.48	2.46	0.015	1.16	0.49	2.40	0.017
归因	0.26	0.06	4.11	0.000	0.01	0.08	0.17	0.867	0.01	0.08	0.16	0.871
调节效应												
惩罚形式×归因					0.51	0.12	4.28	0.000	0.51	0.12	4.19	0.000
控制变量												
性别									0.13	0.19	0.65	0.513
专业									0.01	0.09	0.08	0.937
R^2	0.12^{***}				0.18^{***}				0.18^{***}			
ΔR^2					0.06^{***}				0.00			

注：$N = 120$，$^{***}p < 0.001$。

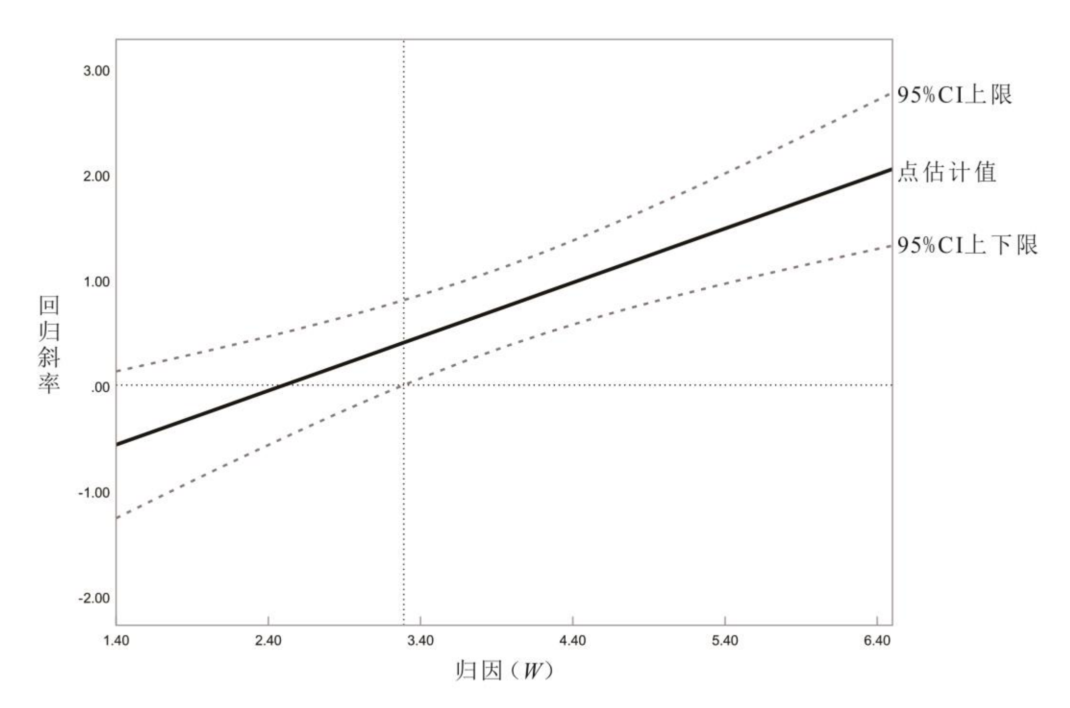


图 8 归因对惩罚形式和能力关系的调节作用

通过图 8 可以看出，当归因大于 3.29 时，惩罚形式影响能力的回归斜率置信区间都在 0 点以上，这说明惩罚形式对能力有显著影响，也就是说，在这种情况下相对于经济惩罚，被试对做出社会惩罚的个体的能力评价更高；而当归因小于 3.29 时，置信区间包含 0 点，此时惩罚形式对能力的影响不显著，被试对做出两种惩罚的个体的能力评价无显著差别。上述结果表明惩罚形式对能力的影响是有条件的：当被试将惩罚归因为集体聚焦时，会对做出社会惩罚的个体更高的能力评价；当被试将惩罚归因为自我聚焦时，对两种惩罚者的能力评价无显著差异。

尽管经济惩罚目前仍然是有关第三方惩罚实验室研究的主流范式，但近年来也有学者开始探讨社会惩罚在合作中的作用机制（Noussair & Tucker, 2005）。Nelissen 和 Mulder（2013）比较了社会惩罚和经济惩罚对合作的促进作用，并发现前者的效果更为显著，而 Wu 等（2016）也注意到，相较于经济惩罚，社会惩罚不仅更有效地促进了个体间的合作，还提升了集体的最终净收益。实验 3 比较了两种惩罚形式对声誉的影响，并发现社会惩罚在一定程度上优于经济惩罚的效应同样体现在了惩罚者的声誉上：当惩罚动机被归因为个体聚焦时，经济惩罚对温暖的负面作用显著高于社会惩罚，而当惩罚动机被归因为集体聚焦时，社会惩罚对能力的正面作用显著高于经济惩罚。先前有关惩罚者声誉的研究多基于经济惩罚（e.g., Barclay, 2006; Hardy & van Vugt, 2006; Kiyonari & Barclay, 2008），本研究的结果意味着引入不同的惩罚形式会为惩罚者的声誉研究提供新的思路。此外，和前两个实验相比，实验 3 的被试并不直接参与博弈，因此不存在角色效应（被试在扮演分配者或接受者时的表现对其

作为惩罚者的声誉的影响），但总体上我们仍然得到了相似的实验结果，这在一定程度上说明本文的结论具有较高的稳健性。

5 总讨论

在二阶社会两难问题中，第三方惩罚本身亦是一种公共物品（Colman, 2006; Hauert et al., 2007），而这种公共物品的提供者——惩罚者——是如何产生并在演化中得到选择是研究者所面对的难题之一。一种较为直观的回答是第三方惩罚能为惩罚者带来积极的声誉以及由此产生的额外收益，从长期来看这部分收益能抵消惩罚成本（Barclay, 2006; Barclay & Kiyonari, 2014），从而为惩罚者带来演化优势。上述理论的前提条件是第三方惩罚所带来的声誉是积极的，但现有文献表明这一预设未必成立（de Kwaadsteniet et al., 2019; Ozono & Watabe, 2012），本文从声誉的双维度、惩罚动机和惩罚形式入手进一步探讨了上述现象的心理机制，从以下方面推进了我们对惩罚者声誉机制的理解。

第一，以往研究倾向于将声誉看作是单维度变量（de Kwaadsteniet et al., 2019），惩罚行为对声誉的影响往往是单向的，要么是正面的（Barclay, 2006; Barclay & Kiyonari, 2014），要么是负面的（Ozono & Watabe, 2012）。本研究基于 Fiske 等（2007）的理论将声誉划分成温暖和能力两个维度，结果表明，惩罚行为对这两个维度的影响方向是相反的，具体而言，惩罚降低惩罚者的温暖评价而提升其能力评价，用中国古典文献的标准来区分，惩罚者似乎更接近智者而非仁者。这意味着，如果未来研究者试图用声誉来解释第三方惩罚的演化机制时，就必须将这两者区分开，换言之，在不同情境下，第三方惩罚会为惩罚者带来截然不同的后果。假设某个群体因遭遇危机而偏好能力突出的成员，那么在这种情况下惩罚者有机会获得更高的权力或社会地位（Gross et al., 2016），因为曾经的惩罚行为导致其具有较高的能力评价；而如果种种原因导致群体更偏好友善温和的成员，那么在这种情况下惩罚者可能将因为惩罚行为而面对不利后果，如遭到排斥或降低得到他人帮助的概率，因为其较低的温暖评价意味着惩罚者不受人喜欢（Geiger & Swim, 2016）。简言之，声誉机制只能部分地解释惩罚者在特定情形下的选择优势，因此必定存在其他有助于惩罚者在演化中得到选择的机制（Dreber et al., 2008）。探讨这些潜在机制是未来研究的重要方向之一。

第二，受经济学和生物学等学科的影响，在有关第三方惩罚的文献中动机在很大程度上被忽视了，而第三方惩罚对惩罚者声誉的影响被简单理解为类似于“刺激-反应”的行为主义模式：惩罚直接引发了他人或正面或负面的评价，而无需考虑驱动惩罚的主观动机。然

而，生活经验和心理学文献指出，人际互动在很大程度上依赖于人们对参与者行为动机的推断（Bigman & Tamir, 2016）。通过引入动机视角，本文证实了动机归因对惩罚者声誉的影响。具体而言，动机被归因为集体聚焦的惩罚能减缓其对温暖的负面作用而进一步提升其对能力的正面影响，相反，动机被归因为自我聚焦的惩罚进一步降低了温暖评价并失去了提升能力评价的积极功能。这一发现意味着在不同的动机归因下，同一惩罚行为对惩罚者的声誉具有截然不同的影响，因此将惩罚动机归因纳入相应的研究中对于理解惩罚者的声誉机制具有重要意义。同时，这一结果也部分地说明了为何声誉机制无法充分解释惩罚者的选择优势，陈思静和杨莎莎（印刷中）对第三方惩罚动机的分析显示，惩罚者的行为在很大程度上是由自利的动机所驱动，而这种自我聚焦的动机反而会阻碍了惩罚者获得良好的声誉。

第三，Guala（2012）指出，真实生活中的第三方惩罚更多地表现为社会惩罚而非实验室环境中的经济惩罚，本研究通过引入惩罚的不同形式并检验其与惩罚动机间的交互作用进一步提高了社会惩罚这一概念的应用范围。先前有研究指出，在促进合作与提高集体收益方面，社会惩罚比经济惩罚更有效（Nelissen & Mulder, 2013; Wu et al, 2016）。本研究的结果显示，这种效应同样存在于惩罚者的声誉中：经济惩罚总体上降低了惩罚者的声誉而社会惩罚总体上提升了惩罚者的声誉，而惩罚动机归因往往扩大了两种惩罚形式对声誉的影响：自我聚焦的归因进一步放大了经济惩罚对温暖的负面作用，而集体聚焦的归因则进一步提高了社会惩罚对能力的正面作用。此外，就两种惩罚形式的交互作用而言，我们发现，当存在社会惩罚这个选项时，经济惩罚总是降低惩罚者的温暖评价而无论个体是否做出社会惩罚，而在能力维度上，单独的经济惩罚可提高惩罚者的声誉，但双管齐下反而对声誉造成了负面影响。这意味着，惩罚作为维护社会规范的一种手段并非多多益善，这一方面呼应了众多学者所提及的惩罚的潜在负面作用（陈思静等, 2020; Fehr & Williams, 2018），另一方面对政策制定者也具有一定的实际参考意义，过量实施惩罚反而可能降低了惩罚者的声誉并损失了社会的运行效率。

第四，本研究的另一理论意义在于初步探索了个体对惩罚动机进行归因的线索。正如Kiyonari 和 Barclay（2008）指出，现实生活中人们很难完整地追踪惩罚行为的前因后果，因此，有理由认为人们总是依据有限的线索来推测惩罚者的动机。本研究的结果初步表明，惩罚者的合作水平在一定程度上起到了线索的作用，即高合作水平意味着该个体的惩罚动机是出于维护集体利益。此外，先前有文献认为社会惩罚通常对合作总是具有积极的促进作用（崔丽莹等, 2017; Nelissen & Mulder, 2013），本研究的结果表明，这可能是因为在先前研究中，不存在其他的信号机制，因此社会惩罚总是被认为出于良善的动机，但当社会

惩罚与其他的信号之间存在明显矛盾时（如惩罚者较低的合作水平），社会惩罚同样会对声誉会造成负面影响，如降低了惩罚者的温暖评价。这使得我们有理由怀疑，在这种情况下，社会惩罚是否对合作依然有积极的促进作用。该发现的意义在于，惩罚行为对声誉或合作的影响并不是存在真空之中，恰恰相反，它根植于惩罚者的种种行为，包括惩罚行为和非惩罚行为，而在后一类型中，某些行为（如合作）会被其他个体当作推断惩罚动机的信号，而某些行为（如作为接受者的表现）却缺乏这种功能。在以往文献中，我们注意到大部分实验室研究为了得出更为明确的因果关系往往人为消除了这些线索，本研究的结果表明通过类似方式得到的结论可能存在一定的片面性。未来研究可从两个方面来改进研究设计：1）将相应的信号线索纳入到研究中，并考察这些线索对惩罚的影响；2）检验其他更多的信号线索以及不同线索间的交互作用。

最后，尽管本文得到了若干有意义的结果，但作为一个探讨惩罚动机、惩罚形式与惩罚者声誉不同维度间关系的探索性研究，本文尚有种种不足之处。首先，就惩罚形式而言，一个值得进一步讨论的问题是经济惩罚与社会惩罚之间的换算关系，换言之，多少单位经济惩罚的强度可被认为等同于相应单位的社会惩罚。如能解决上述问题，那么我们可在控制惩罚强度的基础上进一步比较两种惩罚形式的影响，这样做无疑会极大提高研究结论的说服力。其次，本文虽然初步探讨了合作作为归因线索的作用，但显而易见的是，现实生活中暗示惩罚动机的线索肯定要丰富得多，因此，除了惩罚者的合作水平外，我们尚需进一步探索其他线索以及不同线索间的交互作用，但限于研究技术与文章篇幅，我们未能对此做出进一步的分析。最后，在实验 1 和 2 中，评价者直接参与了与惩罚者的互动，而在实验 3 中评价者并未与惩罚者有直接互动，只是作为旁观者参与了博弈，这分别对应了现实生活中的两种典型情境（即评价者是否直接参与了他/她所评价的事件），但我们未能直接比较两种条件对惩罚者声誉的影响，未来研究可进一步探讨评价者是否参与博弈的影响，这一点具有重要的现实意义，因为当我们在真实生活中扮演评价者时，我们有可能直接参与了该事件，更有可能只是一个旁观者。

参考文献

Balliet, D., Mulder, L. B., & van Lange, P. A. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–630.

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344.

Barclay, P., & Kiyonari, T. (2014). Why sanction? Functional causes of punishment and reward. In P. A. van Lange, B. Rockenbach, & T. Yamagishi (Eds.), *Reward and punishment in social dilemmas* (pp. 182–196). Oxford, England: Oxford University Press.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(4), 936–941.

Beersma, B., & van Kleef, G. A. (2011). How the grapevine keeps you in line: Gossip increases contributions to the group. *Social Psychological and Personality Science*, *2*, 642–649.

Bicchieri, C., Dimant, E., & Xiao, E. T. (2018). *Deviant or wrong? The effects of norm information on the efficacy of punishment* (PPE Working Papers 0016). Philadelphia, PA: Philosophy, Politics and Economics of University of Pennsylvania.

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*(12), 1654–1669.

Bornstein, G., & Weisel, O. (2010). Punishment, cooperation, and cheater detection in "noisy" social exchange. *Games*, *1*(1), 18–33.

Carpenter, J., Bowles, S., Gintis, H., & Hwang, S. H. (2009). Strong reciprocity and team production: Theory and evidence. *Journal of Economic Behavior & Organization*, *71*(2), 221–232.

Chen, H., Zeng, Z., & Ma, J. (2020). The source of punishment matters: Third-party punishment restrains observers from selfish behaviors better than does second-party punishment by shaping norm perceptions. *PloS One*, *15*(3), e0229510.

Chen, S. J., Hu, H. M., & Yang, S. S. (2020). Payment vs. retaliation: Impact of cost form on third-party punishment. *Journal of Psychological Science*, *43*(2), 416–422.

[陈思静, 胡华敏, 杨莎莎. (2020). 支付与报复：成本形式对第三方惩罚的影响. *心理科学*, *43*(2), 416–422.]

Chen, S. J., & Yang, S. S. (in press). Motives of altruistic punishment. *Advances in Psychological Science*.

[陈思静, 杨莎莎. (印刷中). 利他性惩罚的动机. *心理科学进展*.]

Chen, S. J., & Zhu, Y. (2020). The other face of punishment: Detrimental effects of punishment and destructive punishment. *Journal of Psychological Science*, *43*(4), 911–917.

[陈思静, 朱玥. (2020). 惩罚的另一张面孔: 惩罚的负面作用及破坏性惩罚. *心理科学*, *43*(4), 911–917.]

Chen, X., Zhao, G. X., & Ye, H. S. (2014). The forms and functions of punishment in public-goods dilemmas. *Advances in Psychological Science*, *22*(1), 160–170.

[陈欣, 赵国祥, 叶浩生. (2014). 公共物品困境中惩罚的形式与作用. *心理科学进展*, *22*(1), 160–170.]

Colman, A. M. (2006). The puzzle of cooperation. *Nature*, *440* (7088), 744–745.

Cui, L. Y., He, X., Luo, J. L., Huang, X. J., Cao, W. J., & Chen, X. M. (2017). The effects of moral punishment and relationship punishment on junior middle school students' cooperation behaviors in public goods dilemma. *Acta Psychologica Sinica*, *49*(10), 1322–1333.

[崔丽莹, 何幸, 罗俊龙, 黄晓娇, 曹玮佳, 陈晓梅. (2017). 道德与关系惩罚对初中生公共物品困境中合作行为的影响. *心理学报*, *49*(10), 1322–1333.]

de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., & van Dijk, E. (2019). Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology*, *84*, 103800.

de Kwaadsteniet, E. W., van Dijk, E., Wit, A., de Cremer, D., & de Rooij, M. (2007). Justifying decisions in social dilemmas: Justification pressures and tacit coordination under environmental uncertainty. *Personality and Social Psychology Bulletin*, *33*(12), 1648–1660.

Dreber, A., Rand, D. G., Fudenberg, D., & Nowak, M. A. (2008). Winners don't punish. *Nature*, *452*(7185), 348–351.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*(2), 63–87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–140.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*(6928), 137–140.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour, 2*, 458–468.

Fehr, E., & Williams, T. (2018). *Social norms, endogenous sorting and the culture of cooperation*. (ECON Working Papers 267). Zurich, Switzerland: Department of Economics of University of Zurich.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83.

Geiger, N., & Swim, J. K. (2016). Climate of silence: Pluralistic ignorance as a barrier to climate change discussion. *Journal of Environmental Psychology*, *47*, 79–90.

Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the Leviathan: Voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports*, *6*, 20767.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*(1), 1–15.

Hardy, C. L., & van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin*, *32*(10), 1402–1413.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., & Sigmund, K. (2007). Via freedom to coercion: The emergence of costly punishment. *Science*, *316*(5833), 1905–1907.

Houser, D., Xiao, E. T., McCabe, K., & Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior*, *62*, 509–532.

Hu, T. Y., Li, J., Jia, H., & Xie, X. (2016). Helping others, warming yourself: Altruistic behaviors increase warmth feelings of the ambient environment. *Frontiers in psychology*, *7*, 1349.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476.

Jordan, J. J., & Rand, D. G. (2019). Signaling when no one is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *Journal of Personality and Social Psychology*, *118*(1), 57–88.

Kiyonari, T., & Barclay, P. (2008). Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, *95*(4), 826–842.

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84.

Liu, G. F., & Xin, Z. Q. (2014). The effects of punishment impacting on social trust and cooperation: Controversy and interpretation. *Journal of Shanghai Normal University(Philosophy & Social Sciences Edition)*, *43*(1), 146‒152.

[刘国芳, 辛自强. (2014). 惩罚对信任与合作的影响: 争论与解释. *上海师范大学学报(哲学社会科学版)*, *43*(1), 146‒152.]

Nelissen, R. M., & Mulder, L. B. (2013). What makes a sanction "stick"? The effects of financial and social sanctions on norm compliance. *Social Influence*, *8*(1), 70‒80.

Noussair, C., & Tucker, S. (2005). Combining monetary and social sanctions to promote cooperation. *Economic Inquiry*, *3*(3), 649‒660.

Ouss, A., & Peysakhovich, A. (2015). When punishment doesn't pay: Cold glow and decisions to punish. *The Journal of Law and Economics*, *58*(3), 625‒655.

Ozono, H., & Watabe, M. (2012). Reputational benefit of punishment: Comparison among the punisher, rewarder, and non-sanctioner. *Letters on Evolutionary Behavioral Science*, *3*(2), 21‒24.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, *27*(9), 1192–1206.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, *17*(8), 413‒425.

Santos, M. D., Rankin, D. J., & Wedekind, C. (2010). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1704), 371‒377.

Wu, J., Balliet, D., & van Lange, P. A. (2016). Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific Reports*, *6*, 23919.

Xie, D. J., & Su, Y. J. (2019). The evolutionary and cognitive mechanisms of third-party punishment. *Journal of Psychological Science*, *42*(1), 216‒222.

[谢东杰, 苏彦捷. (2019). 第三方惩罚的演化与认知机制. *心理科学*, *42*(1), 216‒222.]

Xie, X. F., Wang, Y. L., Gu, S. Y., & Li, W. (2017). Is altruism just other-benefiting? A dual pathway model from an evolutionary perspective. *Advances in Psychological Science*, *25*(9), 1441‒1455.

[谢晓非, 王逸璐, 顾思义, 李蔚. (2017). 利他仅仅利他吗?——进化视角的双路径模型. *心理科学进展*, *25*(9), 1441‒1455.]

Warmth and Competence: Impact of Third-party Punishment on Punishers' Reputation

CHEN Sijing , XU Yechao

(*School of Economics and Management*, *Zhejiang University of Science and Technology*, *Hangzhou*, 310023, *China*)

Abstract

Third-party punishment (TPP) provides a theoretical explanation to the extensive cooperation among genetically unrelated individuals but also raises a second-order free-riding problem. To solve this challenge, some researchers have proposed the reputational benefits of TPP as a potential explanation. That is, given the positive reputation derived from the punishment, the probability that the punisher would be helped in the future is enhanced. However, a growing body of literature has suggested that punishers' reputation (PR) is not necessarily positive. Three reasons underlying the contradictory findings regarding PR may exist in the existing literature.

Previous research a) views reputation as a uni-dimensional variable that is simply negative or positive, b) fails to take into account punishment motives, and c) only considers financial punishment and overlooks the existence of other forms of punishment.

A series of experiments were conducted to answer the main research questions raised in the current study: a) Does TPP have different effects on the two dimensions of PR (experiment 1)? b) Does the attribution of punishment motives (APM) significantly affect PR (experiment 2)? c) Do financial and social punishment diverge on the effects on PR (experiments 2 and 3)? All these experiments used the dictator game as the experimental paradigm, in which the dictator received an initial endowment and decided to what extent she/he wanted to split this endowment with the recipient, and the observer (financially or socially) punished the dictator for an offer she/he deemed unfair. In experiments 1 and 2, every participant played each role at the same time in random order. In experiment 3, the participants just observed the game. After being presented with the performance of their partners (experiments 1 and 2) or the actors (experiment 3), the participants were asked to rate their reputation using a seven-point Likert scale consisting of six items and attribute punishment motives using a seven-point Likert scale with self-focused or group-focused anchors.

The results of experiment 1 showed that TPP generally reduced participants' evaluation of PR in the warmth dimension and improved their evaluation in the competence dimension. Moreover, the punishment deemed that group-focused anchors could further improve the positive impact of TPP on competence while ameliorating its negative impact on warmth. The results of experiment 2 revealed that participants relied to a large extent on the cooperation level of punishers to determine whether their punishments were group-focused or self-focused, thus affecting their evaluation of PR. Moreover, with the option of social punishment available, financial punishment significantly reduced the evaluation of PR in terms of warmth regardless of their punishment motives, but the evaluation in terms of competence was quite different. In the absence of social punishment, financial punishment improved the evaluation of PR. Otherwise, financial punishment reduced its evaluation. This finding was partially replicated in experiment 3, which further demonstrated by examining the interaction between APM and punishment forms that the APM moderated the effects of the two types of punishment on PR. When deemed self-focused, the financial punishment reduced participants' evaluation of PR in the warmth dimension significantly more than the social punishment. When viewed as group-focused, financial punishment enhanced participants' evaluation of PR in the competence dimension significantly less than social punishment.

In conclusion, this study found evidence that, when considering the impact of TPP on PR, considering the motives underlying punishment and viewing reputation as a multi-dimensional construct is necessary, implying that the reputational benefits of punishing cannot fully explain the selective advantages of TPP and that other factors must have contributed to the selection and diffusion of TPP in evolution.

Key words third-party punishment; social norm; punishment motive; reputation; financial sanction; social sanction

附录：

检验被试理解实验流程及概念的测试题目（举例）

1. 假设在一轮实验中甲分给乙的代币数是 2，丙没有扣减甲的代币，则该轮结束时三者手中的代币数是：（d）
   a 甲：10，乙：0，丙：0
   b 甲：8，乙：2，丙：0
   c 甲：2，乙：8，丙：5
   d 甲：8，乙：2，丙：5
2. 假设 X 的决策主要关注自己代币的多少，Y 的决策主要关注是否大家都能获得较好收益，则 X 和 Y 分别属于：（a）
   a 个体聚焦；集体聚焦
   b 集体聚焦；个体聚焦
   c 个体聚焦；个体聚焦
   d 集体聚焦；集体聚焦
3. 在个体聚焦-集体聚焦题目中得分越高表示：（c）
   a 越关注个体利益
   b 越关注对方利益
   c 越关注集体利益
   d 越关注第三方利益
4. 对某个陈述评价时打分越低代表：（b）
   a 越同意该陈述
   b 越不同意该陈述
   c 越忽视该陈述
   d 越觉得该陈述不重要
5. 假设在一轮实验中甲分给乙的代币数是 3，丙扣减了甲的代币，则该轮结束时三者手中的代币数是：（b）
   a 甲：1，乙：3，丙：0
   b 甲：1，乙：3，丙：3
   c 甲：7，乙：3，丙：5
   d 甲：1，乙：3，丙：5